International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Structured Dirichlet Smoothing Model for Digital Resource Objects

Wafa' Za'alAlma'aitah, Abdullah ZawawiTalib, MohdAzam Osman

Abstract: ***Typically, the results of digital resource object retrieval are made up of whole documents. The problem is that each document contains a large number of metadata units. These units are located in a single document describing different topics. Therefore, retrieving the entire document means retrieving all these units which may be mostly irrelevant to the user's query. The Dirichlet smoothing model usually calculates the probability of having the query in each document and then displays the related documents basedon the query.It is therefore best to retrieve the nearest and relevant metadata units for the query regardless of which document they belong to.To achieve this, astructured Dirichlet smoothing model is proposed in this paper thatcalculates the likelihood between the query and the metadata units instead of between the query and the whole document. The experiments which were conducted on the cultural heritage CHiC2013 collectionhave shown a statistically significant improvement over the traditional Dirichlet smoothing model.***

Keywords: Digital resource objects, Dirichlet smoothing model, Information retrieval.

I.INTRODUCTION

Dirichlet smoothing (DS) model is a model that is used to re-estimate the zero probability for unseen terms by giving them small values derived from the probabilistic values of the seen terms. DS model consists of two parts namely: Query Likelihood Estimation Model (QLEM) and Dirichlet Prior (DP). QLEM is the basic approach for estimating documents using probabilistic language in which q_i is a query term in query Q and D is a document. Given that query terms are independent and employ a unigram language model, the probability of a query given a document $p(Q|D)$ is calculated by applying the Bayesiantheorem that is the product of query terms' probabilities given a document. The probability estimates can be written as

$$p(Q|D) = \prod_{i=1}^{n} p(q_i|D) \quad [1]$$

$$p(q_i|D) = \frac{f_{qi}, D}{|D|} \quad [2]$$

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Wafa' Za'alAlma'aitah*, School of Computer Sciences, UniversitiSains **Malaysia, Penang, Malaysia.** Department of Basic Sciences, Hashemite University, Jordan
Abdullah ZawawiTalib, School of Computer Sciences, UniversitiSains Malaysia, Penang, Malaysia
Mohd Azam Osman, School of Computer Sciences, UniversitiSains Malaysia, Penang, Malaysia

where,

f_{qi}, D : Number of occurrencesof query term q_i in document D

$|D|$: Length of document D

The DS model (Zhai, 2002)is given as

$$p(q_i|D) = \frac{f_{qi}, D + \mu p(qi|c)}{|D| + \mu} \quad [3]$$

According to Zhai (2002), the optimal prior value for μ is 2000.

Eq. (3) finds the probability between the query and the document, and thus the model retrieves the documents related to the query. In DROs, when retrieving documents as results for a particular query, each document contains a large number of metadata units, and thus the user is required to re-search within each document. Here lies the problem when there are more than one alternative search spaces. In this case, Eq. (3) needs to be modified to handle more search spaces.

DROdoes not only contain documents and collections, it also has levels that can be inside another level (Brocks et al., 2001; Hatano et al., 2002). Consequently, the unit of information to be returned to the users can vary (Mataoui et al., 2015). Furthermore, the smoothing models used such as Dirichlet smoothing model, may increase the number of returned information units (Smucker et al., 2005). In order to enhance the retrieval performance of DROs, the Dirichlet smoothing model has to take into account the document structure (Liu et al., 2008) as in the proposed structured Dirichlet smoothing (SDS) model presented in this paper.

The rest of paper is organized as follows: Section 2 presents the related work. The proposed SDS model is described in Section 3. The experimental results and discussions are presented in Section 4. Finally, Section 5 concludes the paper.

II.RELATED WORK

Dirichlet smoothing makes smoothing dependent on the document size. They are more likely to require less smoothing. If multinomial distribution is used to represent a language model (LM), the conjugate prior of this distribution is the Dirichlet distribution (Zhai, 2002). As μ gets smaller, the collection model also becomes smaller, and more emphasis is given to the relative term weighting. Zhai (2002) and Zhai and Lafferty (2004) have reported that the

optimal prior value for μ is around 2000.In He and Ounis (2005),Dirichlet smoothing is reformulated based on measuring the correlation of the normalized term frequency with the document length for a given query term. Zhai (2008)has shown that long documents are less impacted by μ and should be tuned, or the average document length is picked. Dirichlet smoothing language model is generally considered to be more effective than other smoothing-based language models, especially for short queries (Strohman *et al.*, 2005). Furthermore, Azzopardi and Losada (2007) and Losada and Azzopardi (2008) have demonstrated that the Dirichlet smoothing tends to retrieve many short documents and few long documents.In a recent study by Tan (2015), a new smoothing parameter called α was suggested beside the μ smoothing parameter. The α smoothing parameter refers to the value from term similarity. It can integrate the term links inside the document. It is assumed to increase the document if the term from the query has a link with the term in term of similarity matrix. The experiment was conducted on CHIC2013 collection. The experiments results show that the proposed method has produced a significant improvement on the retrieval performance giving 0.525 and 0.48 for MAP and P@10 respectively. Along this line, Ogawa *et al.* (2016) proposed an extended Dirichlet smoothing model using a dynamic document set obtained from web pages as a dynamic collection. The proposed model combines together the static and dynamic document collections. However, the model adds further smoothing parametercalled V with a value of 80 for the language model built using the dynamic document collection. For static documents, the smoothing μ parameter was used with a value of 230.

According Zhai and Lafferty (2001), the query likelihood model has generalized to the divergence scoring method, by modeling the query separately. Among the many proposed approaches of the language model (LM), the most popular and fundamental one is the query-likelihood language model which was shown to be theoretically superior and this has been confirmed experimentally inBruza and Song (2003), Mei *et al.* (2007), Lv and Zhai (2009) and Hui*et al.*(2011). Cummins *et al.* (2015)proposed the Smoothed Polya Document language model which incorporates word burstiness only into the document model. It uses the Dirichlet compound multinomial to model documents in place of the standard multinomial distribution, and the proposed model uses the standard multinomial to model query generation. The experiments were conducted using the TREC datasets. The experimental results show that the query likelihood model with Dirichlet smoothing can be implemented as effectively as the traditional retrieval.

III.STRUCTURED DIRICHLET SMOOTHING MODEL

The main target for the proposed SDS model is to addresses the whole document retrieval problem(Larsen *et al.*, 2006). The problem is that each document contains a large number of metadata units. These units are contained in a single document containing different topics. Therefore, retrieving the entire document means retrieving all these units which may be mostly irrelevant to the user's query. The DS model usually calculates the probability of having the query in each document and then displays the related documents to the user's query.It is therefore best to retrieve the nearest and relevant metadata units for the query regardless of which document they belong to.To achieve this, the Dirichlet smoothing model needs to calculate the likelihood between the query and the metadata units instead of between the query and the whole document. Let the collection be denoted as C, document as D, metadata unit as U, and query as Q. Eq. (3) will be reformulated as

$$p(q_i|U) = \frac{f_{qi},U + \mu p(qi|D) + \mu p(qi|c)}{|U| + \mu} \quad [4]$$

where,

U : Metadata unit

f_{qi},U : Number of occurrences of the query term q_i in metadata unit U

μ : Smoothing parameter (equal to 2000)

$|U|$: Length of metadata unit

The principle work of the proposed equation is very similar to the principle of the original Eq. (3). However, the proposed equation gives the first priority to searchingin the metadata units.If the probability of the query term given metadata, $p(q_i|U)$ is equal to 0 then the searchgo to the documents, and if probability of the query term given document, $p(qi|D)$ is equal to 0 then the searchis turned to the whole collection to find the probability of the query term in the collection, $p(qi|c)$. This is done in order to avoid the possibility of zero probability in one of the constituent terms of the query which gives the probability of zero for the entire query although some query terms may have a non-zero probability.

It can be observed that the more the search space is, the more likely the presence of the terms will be. Thus, this reduces the probability with zero value and increases the chance of finding unseen terms. In addition, it retrieves the related metadata units and grouped them together from all the documents to answer a user's query. Note that the value of μ parameter in the SDS model is 2000 regardless of the length of the search areas and their contributions mass in the missing query terms.

IV.EXPERIMENT AND RESULTS

This research used CHIC2013 English collection. We choose this collection because it is the only available collection in digital resources with testing query based on documents and metadata unit. The collection provides 1107 documents and 17 testing queries regarding the metadata units. The model was evaluated over the 17 queries because no additional queries are available with their relevant metadata units in the CHiC2013 collectiontesting queries

(1.8 terms per query, on average). The testing queries represent the relevance of each document and metadata unit for each query. For all of our experiments, we report the performance of each model using three standard measures: Mean Average Precision MAP, Precision at top 10 documents p@10 and precision recall curve (Manning, 2013). Two-tailed paired t-test was used to test whether the differences between performances of the models are statistically significant. Two benchmarks were used.The first is the basic language model with Dirichlet smoothing model (DS) as defined by (Lafferty & Zhai, 2001). The second benchmark is the extended Dirichlet smoothing with increased document size (EDS+) proposed by Tan(2015).

The results for the EDS, DS and EDS+ models are shown in Table 1 in terms of MAP and P@10 measure. From the table, we notice that the MAP for EDS model compared with the MAP for DS model gives an improvement of 8.10%, and the MAP for EDS model compared with the MAP for EDS+ model gives an improvement of3.70%. Furthermore, it is necessary to highlight that the EDS model compared with DS model and EDS+ model achieves improvements of 8.10% and 4.50% respectively of based on P@10. In addition, Figure 1 shows that the Precision-Recall curve for DS model and EDS model. Based on this figure, the precision of EDS model gains higher than DS model at different recall points, and it shows that the EDS model helps to improve the retrieval results.

Table.1 MAP and P@10 for DS, EDS+ and SDS model (* indicates the best performance)

Model	MAP	P@10
DS	0.501	0.412
EDS+	0.545	0.48
SDS model	**0.582***	**0.525***
Improvement(SDS, DS)	8.100%	8.10%
Improvement(SDS, EDS+)	3.70%	4.50%

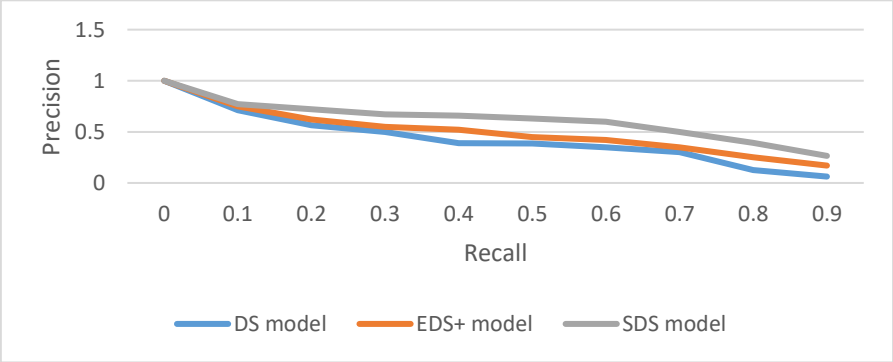

Fig.1 Comparison of SDS with DS and EDS+ by using averaged 9-point precision recall curve

V.CONCLUSION

In this paper, a SDSmodel has been proposed to improve the DROsretrieval. The aim of the proposed SDS model is to addresses the whole document retrieval problemwhich affects negatively onthe performance of DRO retrieval. The proposed SDS model improves the likelihoodequation to calculate the likelihood between the query and the metadata units instead of between the query and the whole document. The SDS model is able to retrieve the metadata unit which leads to improvement of the performance of DRO retrieval. However,the Dirichlet smoothing model uses a fixed value of 2000 for the μ smoothing parameter. Furthermore, the μ parameter plays a strong role in finding the value of unseen terms as a contribution to avoid the zero-probability value. The fixed value of the μ parameter becomes inappropriate and needs to be automatically estimated for structured documents. Hence, for future work, the performance of DROs retrieval can be improved by enhancing μ parameter in theDS model to avoid the zero-probability value which leads to a decrease in the DRO retrieval performance.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of UniversitiSains Malaysia for this research under the Research University Grant entitled “A Two-Stage Expansion Method for Retrieval of Metadata Content in Cultural Heritage Collection. ”

REFERENCES

1. Azzopardi, L., andD.E.Losada. 2007. Fairly retrieving documents of all lengths. In Proceedings of the First International Conference in Theory of Information Retrieval (ICTIR 2007), pp. 65-76.
2. Berger, A., and J. Lafferty. 1999.Information retrieval as statistical translation. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 222-229. ACM

3. Brocks, H., U. Thiel, A. Stein, A. and A. Dirsch-Weigand. 2001. Customizable retrieval functions based on user tasks in the cultural heritage domain. In International Conference on Theory and Practice of Digital Libraries, pp. 37-48. Springer
4. Bruza, P., and D. Song. 2003. A comparison of various approaches for using probabilistic dependencies in language modeling. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 419-420. ACM
5. Cummins, R., J.H. Paik, and Y. Lv.. 2015. A Pólya urn document language model for improved information retrieval. ACM Transactions on Information Systems (TOIS), *33*(4): 21.
6. Hatano, K., H. Kinutani, M. Yoshikawa andS. Uemura. 2002. Information retrieval system for XML documents. In International Conference on Database and Expert Systems Applications, pp. 758-767. Springer
7. He, B., andI. Ounis. (2005). A study of the dirichlet priors for term frequency normalisation. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 465-471. ACM
8. Lafferty, J., and C. Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 111-119. ACM
9. Laitang, C., K. Pinel-Sauvagnat and M. Boughanem. 2013. Estimating structural relevance of XML elements through language model. In Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, pp. 41-46.
10. Larsen, B., A. Tombros, and S. Malik. 2006. Is XML retrieval meaningful to users?: searcher preferences for full documents vs. elements. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 663-664. ACM
11. Liu, S., C.A. McMahon and S.J. Culley. 2008. A review of structured document retrieval (SDR) technology to improve information access performance in engineering document management. Computers in Industry, 59(1): 3-16.
12. Losada, D.E,. and L. Azzopardi. 2008. An analysis on document length retrieval trends in language modeling smoothing. Information Retrieval, 11(2): 109-138.
13. Lv, Y., and C. Zhai. 2009. Positional language models for information retrieval. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 299-306. ACM
14. Manning, P. (2013). Introduction Drugs and Popular Culture (pp. 10-13): Willan.
15. Mataoui, M.H., F. Sebbak, F. Benhammadi andK.B. Bey. 2015. Query Expansion in XML Information Retrieval: a new Approach for terms selection. In Modeling, Simulation, and Applied Optimization (ICMSAO), 2015 6th International Conference on, pp. 1-4. IEEE
16. Mei, Q., X. Ling.M. Wondra, H. Su and C. Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In Proceedings of the 16th international conference on World Wide Web, pp. 171-180. ACM
17. Nallapati, R., and J. Allan. 2002. Capturing term dependencies using a language model based on sentence trees. In Proceedings of the eleventh international conference on Information and knowledge management, pp. 383-390. ACM
18. Ogawa, K., T. Murahashi, H. Taguchi, K. Nakajima, M. Takehara, S. Tamura, and S. Hayamizu. 2016. Spoken Document Retrieval Using Neighboring Documents and Extended Language Models for Query Likelihood Model. In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, pp.186-190.
19. Ogilvie, P., and J. Callan. 2003. Language Models and Structured Document Retrieval.
20. Ponte, J. M., and W.B. Croft. 1998. A language modeling approach to information retrieval. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275-281. ACM
21. Smucker, M.D., D. Kulp and J. Allan. 2005. Dirichlet mixtures for query estimation in information retrieval. Center for Intelligent Information Retrieval.
22. Strohman, T., D. Metzler, H. Turtle, and W.B. Croft. 2005. Indri: A language model-based search engine for complex queries. In Proceedings of the International Conference on Intelligent Analysis, pp. 2-6. Citeseer
23. Tan. 2015. Extended language model in cultural heritage collection(PhD Thesis), Universiti Sains Malaysia.
24. Zhai. C. 2008. Statistical language models for information retrieval. Synthesis Lectures on Human Language Technologies, 1(1): 1-141.
25. Zhai, C., andJ. Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS), 22(2): 179-214.
26. Zhai, C. (2002). Risk minimization and language modeling in text retrieval (PhD Thesis), Carnegie Mellon University.
27. Zhai, C., andJ. Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In Proceedings of the tenth international conference on Information and knowledge management, pp. 403-410. ACM